汉字的熵与语言形式化*

王　　迈*

〈Summary〉

Characters are the second carrier of language following voice the first. Chinese characters have a high entropy system since their features like form complexity, enormous amount, clear distinction and big load of information can be explained sound from the aspect of information entropy. Man has an extremely limited instant perception of information sequence digits but has an extraordinarily powerful instant cognitive ability of the information of each sequence. Therefore, the best information sequence for learning and memorizing that has fewer digits with more information load, which is the characteristic of high entropy system and of Chinese characters system as well. Natural language coding (formalization) should follow C. E. Shannon channel coding theory and the uniqueness principle. The relation between information entropy and coding digits can be seen as the distinctions of different numerical systems. Information entropy is one of the standards of examining commonness and individuality among languages and of measuring how far a language has developed, and reflects the complexity of the concept system expressed by the lexical system of a language and can thus be called concept entropy.

1．文字的性质与语言形式化

文字是语言的书面表达形式，是语言发展到高级阶段的产物。文字的出现极大地扩展了话语信息传递的时间和空间，促进了人类思维成果的积累和传承。

语言的发展经历了两次重要的由动机驱动的事件。其一是表达、交流的动机，它直接促成了人们选择声音这个样态丰富又易于调节的媒介作为思维信息的载体，语言因此而诞生。其二是记录、传递的动机。有声语言在数万年的发展过程中生产出日渐丰富的思维成果，而声音转瞬即逝的特点使其无法将话语信息固化下来，造成思维成果的大量流失，制约了人类文明进步的速度。一种状态稳定、不受时间和空间限制的新媒介成为语言和社会加快发展的迫切要求。线条化的图形具有可视、信息量大、易分辨的特点，当它被记录在某类物质上，它就同时具有了承载物的物理特性，例如金属的坚硬、石头的不朽、竹片的轻薄或者纸张的柔韧等等，这些都是话语信息得以长时间保留或者长距离传递不可或缺的条件。线条化的图形弥补了声音在可视性和时空延展性上的不足，自然地成为思维信息继声音之后的第二载体，称为文字。

虽然文字可以承载思维信息，但它并不与思维直接相关，而是以语言作为中介。换言之，文字是语言的符号，语言是思维的载体（图表 1）。思维作为大脑的内部活动，不受外部因素的直接控制，

我们只能通过刺激间接地影响或干扰思维；语言是思维得以表述和传递的有形载体，具有社会约定俗成的性质，可以接受有限的人为干预，例如规范化；文字是语言单位的形式化，虽然受到具体语言的影响而个性鲜明，但本身具有较强的独立性，因此文字可以改革、创新甚至废止。

图表 1　思维、语言和文字的关系

按照文字所表达的语言单位的不同，可以将其划分为语段文字、词符词素文字、音节文字、音素文字四类。语段文字是最原始的文字类型，它表示一个话语片段，叙述的通常是一个场景或一起事件。语段文字的字形复杂并具有高度的形象性，但由于其尚未和语言单位建立起严格的对应关系，字形也没有固定下来，常被看作是由图画过渡到文字的中间阶段，称为文字画。词符文字表示一个词，相当于一个概念或概念间的关系，如古汉字；词素文字表示一个词素，是音义结合的最小语言单位，如现代汉字。词符词素文字具有数量庞大、字形复杂的特点，经过分解可以得到数量较少、结构简单的字形构件；字形构件不表示语言单位，且变体丰富、位置不定，因此不是文字单位。音节文字表示一个音节，音节是最自然的语音单位，数量通常在几十到上千，音节文字的数量远少于词符词素文字，例如日语假名的数量在一百个不到。音素文字表示的是最小的语音单位——音素或音位，它可以灵活地组成音节，因此数量更少，如拉丁字母和斯拉夫字母都在三十个左右。

如果把语音作为语言本身的形式，那么文字的诞生就是语言发展到一定阶段自发形成的进一步的形式化过程。语言的既有形式是自然形成的，是语言经历了上万年发展变化的结果，也是纷繁的言语变体通过长期竞争优胜劣汰的结果。因此，发展至今的各种语言都是一个相对完善的体系，可以高效满足信息的传统传递方式的需要。但是，随着社会的发展，各种现代化的信息处理和传送方式层出不穷，作为信息载体的语言，不再局限于传统使用领域，它必须自我完善以适应各种现代化应用的需要。然而，语言固有的形式并不能很好地满足这一需求，日益表现出其局限性。因此，我们有必要使用各种物质形式分析并重塑自然语言，拓展自然语言的适应性，这是语言形式化研究的目标之一。我们首先探讨与形式化密切相关的语言文字单位的信息量。

2．熵与语言单位的信息量

熵（Entropy）原是热力学概念，指系统中微观粒子运动的无序程度。熵同系统信息量的大小密切相关：系统越混乱无序，不确定性就越高，可能存在的状态就越多，容纳的信息量就越大。例如，世界杯足球赛中，32 支球队有 32 种可能的冠军结果，不确定性为 32 取 1，从电台获得冠军消息，就可以消除这些不确定状态，假设每支球队夺冠的概率相等，我们就说冠军消息具有 $H_1=\log_2 32=5$bit 的信息量；同样的情况，如果是四国女排邀请赛，冠军消息就只能消除 4 种不确定

状态，信息量为 $H_2=\log_2 4=2$bit。现代信息论之父 C.E.Shannon 最早注意到系统的无序程度同信息量的关系，他把熵的概念引入信息论，作为信息量的量度，称为信息熵。

语言也可以从信息量的角度加以考察。语言单位是语言系统的“微观粒子”，在语言规则的制约下，语言单位的隐现具有一定的自由度（无序性）。言语交际中，每一个语言单位的出现都能够一定程度地降低话语的不确定性，这体现了它在话语中的交际价值。我们经常称赞一段讲话“言简意赅”、“内容丰富”，或批评其“老生常谈”、“废话连篇”，其实都是对话语信息量做出的感性评价。现在，有了熵的概念和概率工具，我们就可以对语言单位的信息量做出定性和定量的分析。

信息熵使我们更深刻地认识语言，语言的个性特征以及语言间的差异，很多都可以从信息熵的角度得到解释；一些争论不休的问题，也在语言单位的信息熵中找到了答案。

3．汉字的熵

汉字有着超过 3500 年不间断的发展历史，是迄今世界上仍在使用的唯一的语素文字，用汉字写成的文献典籍浩如烟海，记录并传承着汉民族思想文化的结晶。然而，这样一套杰出的文字系统，自二十世纪上半叶起，其存在的合理性遭到了广泛质疑，有人甚至提出了“废止汉字”、“汉字拉丁化”的口号。

汉字遭质疑的缺陷主要有：

① 字形复杂，难学难用。拉丁字母只有 1～2 画，汉字的平均笔画在 8～10 画。

② 数量庞大。英语字母为 26 个，俄语 33 个，汉字的数量达到几千～上万。

③ 表音能力差。即使是形声字，字音偏离声旁的情况也占三分之一强。

④ 在计算机上编码困难、录入困难、显示困难，不能适应信息化时代的需要。

到了八十年代，信息熵理论的引入使汉字的诸多特点得到了合理解释，有力地驳斥了汉字缺陷论。冯志伟（1984）对容量 12370 字的语料库进行统计，测出汉字的熵为 9.65bit，同时指出：当汉字容量不大时，包含在一个汉字中的熵随着汉字容量的增加而增加，当汉字容量达到 12366 个字时，包含在一个汉字中的熵就不再增加了，这就是著名的“汉字容量极限定律”。

语种	汉语	俄语	德语	英语	西班牙语	意大利语	法语
文字的熵（bit）	9.65	4.35	4.10	4.03	4.01	4.00	3.98

图表 2　汉字与音素文字的熵值比较

与汉字相比，音素文字的信息熵要小得多，大致在 4bit 左右，图表 2 列出了印欧诸语言的字母熵。我们发现，正是汉字与众不同的高熵值，造成了被人们质疑的诸多缺陷；但是，事物都具有两面性，高信息熵也给汉字带来了特别的优势：

① 区别度高。与拉丁字母的一维线性排列不同，汉字有左右、上下、内外等结构，是二维图形，因此区别度高，容易辨识。

② 信息承载量大，占用空间则相对较小。现代汉字是语素文字，对应于一个概念或构词理据；英语语素大多需要一组字母表示。同一文件的多国译文中，汉译本总是最薄的。汉字还承载着丰富的美学信息，形成了汉字特有的书法艺术。

③ 字音与字义直接关联，阅读速度快。音素文字表达的是音位层次的区别性意义；汉字承载的意义达到了概念级别，可以直接作用于人脑的意义感知区域，所以汉语默读可以达到很高的速度，所谓“一目十行”。

④ 如前述，汉字数量众多并且表达语言单位的语义级别较高，这使汉字作为输入单位彼此可以建立丰富的意义关联，在词的内部通过音或形的结合有效降低重码率，新型的输入法已经达到不低于英语的录入速度。此外，计算机海量存储设备和高分辨率显示设备的发展也使字库存储和显示不再成为问题。

人类获取和存储信息的方式同自身的生理和心理特点密切相关。对于一个线性排列的信息序列，例如一串数字或一个言语片段，其信息总量主要取决于：①序列的位数；②每一位所承载的信息量。实验证明，人类瞬间同时记忆的离散的符号可以达到约 5～7 位，即使经过长期的严格训练，最多也只能同时记忆十多个符号，这是人类生理和心理的极限。但是人类对符号的辨识能力却几乎是无穷的，正常人在经过几年的学习后很容易就可以在瞬间分辨出几千个汉字中的任何一个。我们得出的结论是：人类对信息序列的位数的瞬间感知能力极其有限，而对信息序列每一位所包含的信息的瞬间区别能力却非常强大。这一显著倾向告诉我们，适合人类学习记忆的信息序列类型是：①位数较少；②每一位所蕴含的信息量较多。而这，正是高熵系统的特点，也是汉字体系的特点。

可见，对汉字的评价应该充分考虑高熵系统的特点，以及人类自身的生理、心理特点。我们完全有理由相信，汉字是最适合表达汉语的文字体系，也是符合现代信息处理要求的文字体系，它不仅不比音素文字“劣等”，反而体现出一定程度的优越性。

4．熵与形式化

C.E. Shannon 给出了“仙农信道编码定理”：在一个非扩展的无记忆信源中，用二进制代码表示的码字的长度不能小于信源的熵。这是说，如果用二进制表达汉字系统，每个汉字分配的空间不能小于汉字的平均信息熵 9.65bit。但这只是一个理想的极限值，真实系统必然存在一定的冗余度，因此，现代计算机一般使用 16bit（2Byte）编码一个汉字，并且，当字库的容量超过 65536 字时，16bit 仍然是不够的。

信息熵同编码位数的关系可以形象地描述为不同数制间的特征差异。数制即数的表示方法，它用一组固定的数字符号和一套统一的规则来表示数。人们比较熟悉的十进制以十为基数，使用 0～9 十个数字符号，记录规则“逢十进一”。推而广之，N 进制以 N 为基数，使用 N 个数字符号，逢 N 进一。其中 N 是自然数，例如二进制、八进制、十六进制等。

数制是信息熵的理想数学模型，它摒弃了所有可能的冗余信息，并且保证随机事件的概率都相等。通过数制的比较，很容易获得高熵系统和低熵系统的特点差异（特别是熵的大小同编码位数的

关系）。图表 3（a）是数量“9”在二进制和十进制中的表示，表（b）是概念“火”在英语和汉语中的表示，两者的差异有相似的地方。

	二进制	十进制
表达形式	1001	9
位数	4 位	1 位
每一位的信息量（区别度）	低（区别于另外 1 个数字）	高（区别于另外 9 个数字）
不确定性（基元数）	2 个（0 和 1）	10 个（0～9）
信息熵（bit）	$Log_2 2 = 1$	$1og_2 10 \approx 3.32$

（a）“9”在二进制和十进制中的特征差异

	英语	汉语
表达形式	Fire	火
位数	4 位	1 位
每一位的信息量（区别度）	低（区别于另外 25 个字母）	高（区别于数千汉字）
不确定性（文字数量）	26 个（a ～ z）	几千个（常用）
文字熵（bit）	4.03	9.65

（b）“火”在英语和汉语中的编码特征差异

图表 3　高熵系统与低熵系统的特征差异

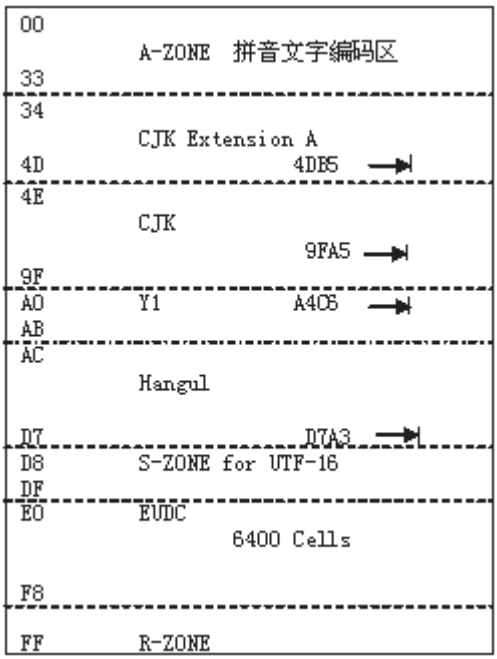
如果用二进制来编码一个十进制符号（阿拉伯数字），那么给它分配 3.32bit 就是最经济的，少于 3.32bit，就会造成溢出；多于 3.32bit，就会产生冗余。语言单位的编码位数则没有理想模型那样精确，因为语言系统为了保证话语信息的稳定性和可靠性，包含了大量的冗余信息，它们要占据可观的编码空间，但总的来说，语言单位编码长度与它的信息熵成正比。

字符编码要遵循的另一个原则是唯一性原则，即在一个确定的编码空间内，每一个二进制码字获得的字符解释必须唯一。它与仙农信道编码定理共同保证了一个字符能够获得计算机充足并且专一的支持。

此外，好的编码方案还应具备如下特点：①系统性。编码地图能够反映文字系统的构成，其布局有规可循，具有模块化特征。②前向兼容性。保证历史文档的可持续利用。③后向可扩展性。为字符集的扩充预留空间，同时支持后续版本的前向兼容性。④灵活性。在同一编码方案下，可以实现灵活多样的转换格式，以适应不同的需求（例如 Unicode 下的 UTF7、UTF8、UTF16 等子方案）。⑤稳健性。也称鲁棒性（Robustness），当传输错误、数据损坏等极端情况发生时，字符集能够把

影响限制在最小的空间内（例如本字符内），而不影响后续字符的正确解释。

1993 年发布的 ISO.10646 定义了一种通用字符集（Universal Multiple-Octet Coded Character Set，UCS 或 Unicode），它是世界上所有已知字符集的超集，并为将来可能出现的字符预留了足够的空间。UCS 规定了一个 4 字节（31 位）的编码空间，并按字节由高至低分别定义为 Group、Plane、Row 和 Cell 四个级别，理论上可以存储 2^{31} 个字符，如此庞大的空间几乎是不可能被填满的，迄今被使用的也只有 0x0000～0xFFFD 共 65534 个码位，相当于第一个 Plane，称为基本多文种平面（Basic Multilingual Plane，BMP）。BMP 是 UCS 的一个子集，称为 UCS-2，Unicode 通常就是指对这个子集的实现，但仅仅是这一个子集，就几乎涵盖了世界上所有文字的主要字符。BMP 的空间规划如图表 4 所示：



图表 4　基本多文种平面的字符构成

其中，大部分拼音文字（包括 ISO.8859 中的大部分字符和日语假名等）位于 A-Zone 区；大部分汉字(包括日、韩、越等国使用的)位于 CJK 区;增补的扩展汉字集位于 CJK ExtensionA 和 0xFFFF 后的部分空间；韩语文字位于 Hangul 区；YI 区包含彝、八思巴、爪哇等一些使用人口不多的文字；S-Zone 为 UTF16 转换方案专用区；EUDC 为保留私用区，用户可根据需要增加自创文字；R-Zone 为限制使用区，包含一些特殊字符和兼容字符等。

UCS 给每个字符分配了唯一的代码，使得现存各类字符集中的任何字符都可以在 UCS 中找到映射。例如，汉语“上外”的 Unicode 是 U+4E0A U+5916；日语“はな”是 U+306F U+306A；韩国语“학교”是 U+D559 U+AD50；其他文字类同，这样，UCS 就实现了世界文字符号的大一

统，“在一个文档中同时显示多国语言”的理想已经成为现实。

UCS 为我们提供了字符编码方案，但并未规定具体的实施细则。在英语国家，传输纯粹的英语字符只需要 8 位，如果仍旧用 2 个字节编码，高字节就得全部置 0，这是十分浪费的，由此在 UCS 基础上制订了 UTF-8（UCS Transformation Format 8），这是一种传输标准，是对 UCS 编码的具体实现。同样，我们可以根据实际需要制定多种实施标准，除了 UTF-8，常见的还有 UTF-7/UTF-16/UTF-32 等。以 UTF-8 为例，它采用变长技术，用 1 个字节传输 UCS 中的 ASCII 字符；用 3 个字节传输 UCS 的 2 字节字符（例如汉字），具体实现原理如下：

0000 0000～0000 007F | 0xxxxxxx

0000 0000～0080 07FF | 110xxxxx 10xxxxxx

0000 0800～0000 FFFF | 1110xxxx 10xxxxxx 10xxxxxx

0001 0000～0010 FFFF | 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

仍以“上外”为例，U+4E0A U+5916 的二进制串分别为 100111000001010 和 101100100010110，将其嵌入模板第三行，得到 111001001011100010001010 和 111001011010010010010110，UTF-8 的编码即为%E4%B8%8A 和%E5%A4%96。同理，如果是拉丁字母，只需嵌入模板第一行，采用一个字节就可以完成传输。

除此之外，还牵涉到一些表达细节，比如 Mac 机和 PC 机对 Unicode 双字节的顺序解释正好相反，例如“上”有时是 4E 0A 有时是 0A 4E，为了确定顺序，可以在文件头加入 BOM（Byte Order Mark）。UCS 中有一个特殊的 0xFEFF 字符，称为“无间隔不中断符”，而字节相反的 0xFFFE 在 UCS 中不存在，把它加在文件头，就可以据此判断字节顺序：如果是 FEFF，就称为 Big Endian；如果是 FFFE，就称为 Little Endian。

UTF-8 不存在字符顺序问题，但凭借 BOM 可以判定编码方式。0xFFFE 的 UTF-8 编码是 EF BB BF，如果文件开头出现这样的字串，可以确定其为 UTF-8 编码。

目前 Windows 的内核已经改用 Unicode 编码，这就做到了从底层支持世界上所有语言共用一个平台。但是，现存的大量文档和软件还是用特定字符集编写的，为了向前保持兼容，Windows 提供了代码页（Code Page）机制，这是一种折中方案，用户根据所处地区设定系统的缺省代码页，系统则根据代码页将软件或文档中的 ANSI 字符自动映射为 Unicode 字符。仍以“上外”为例，其在地方软件中的内码为 0xC9CF 0xCDE2，Unicode 编码为 U+4E0A U+5916，代码页其实就是这两种编码的映射关系表。常见的代码页有：CP936 对应 GBK； CP950 对应 Big5；CP932 对应日语 Shift-JIS；等等。随着 Unicode 的普及，代码页最终将被取消。

5．隐藏在信息熵下的语言共性

语言没有优劣之分，谈论语言的优劣，如同谈论种族的优劣，最终会陷入民族沙文主义的泥沼。但是语言有发达程度的差异，语言的发展总是遵循词汇量增加、语法复杂化、文字从无到有的过程。为了比较两种语言的发达程度，我们需要一个客观的评判标准。

有这样一些可能的标准：①语言使用的广泛度（包括人口数量和国家数量）；②词汇量；③语法的复杂度（包括语法规则的数量和表达能力）；④文字的复杂度（包括字形的复杂度和文字的数量）；⑤是否适应信息化处理手段的需要（如计算机等）；⑥该语言生产的文献的数量；⑦语言文字的优美度和艺术感染力；等等。这些标准都有一定的参考价值，但都没有找到问题的关键。

随着信息熵概念的提出，从上述第五条标准又派生出这样的观点：汉语是高熵语言，高熵语言是落后的语言，阻碍了汉语数字化的进程，它必将被类似英语这样的低熵语言所取代。持这种观点的人不在少数，但它却存在基本原理错误：汉语的高熵指的是汉字，而文字只是有声语言的书面表达形式，是第二性的，不足以代表语言的全部；汉语的其他要素，如语法、语音、词汇等，熵值的高低还有待测量；即便如此，信息熵能否作为语言发达程度的标准，还是值得商榷的。

但是，上述观点也带给我们一个启示：考察和语言相关的各种要素的信息熵（或复杂度），如果能够找到与语言发展相应的普遍发展轨迹，或者语言之间同一要素信息熵的一致性，那么该要素很可能就是判断语言发达程度的正确标准。我们假设汉语、英语、日语是发达程度相近的语言，对其做如下考察：

1）语音。音节是最自然的语音单位，由多寡不同的一组音素构成，音素数量越多，可能出现的排列形式就越多，蕴含的信息量就越大，对语义的区别作用就越显著，音节熵值就越高；反之亦然。

日语的音节结构简单，基本音节由 1～2 个音素构成，拗音音节由 3 个音素构成（中间固定为[i]）。日语的音节数量较少，大致为 100-120 个（有争议）。音节结构简单意味着承载的信息量少，构成一个概念（词）所需的音节数就多，在日语词汇中，三音节以上的词占了绝大多数。

汉语的音节结构比日语复杂。它由声母、韵头、韵腹和韵尾四部分构成（允许复元音，不允许复辅音），加上超音段的声调，可以同时出现 5 个音位。因此汉语的音节数大大超出了日语的数量，达到了 1200 个左右。汉语音节信息承载量大，构成词的音节数较少，现代汉语中，双音节词占优势。

英语的音节结构更复杂。除了允许复元音，也允许复辅音，甚至是三合或四合复辅音，例如 glimpsed [glimpst]，由 7 个音位构成（虽然这是一种极端情况）。英语的音节数超过了汉语，达到 1600 个左右。在英语中，常用词以单音节和双音节占优势。

考察结论是：三种语言的音节熵为英语 > 汉语 > 日语，并且差异较大，显然不能反映语言的发达程度。此外，多数语言的音位都稳定在几十个，语言间差异不大，古今变化也不大，也不能成为评判语言发达程度的标准。

2）语法。对语法规则进行信息熵的估计是困难的，因为规则是人为归纳出来的，抽象程度可高可低。语法所反映的主要是词与词的关系意义，不同语言中归纳出的语法意义不尽相同，有的语法意义有相应的语法形式，有的则没有（零形式）。语法规则的表达手段也各不相同，有的借助词形变化，有的借助词序和虚词的使用，有的借助词后的黏着成分。尽管如此，我们仍然可以凭经验认定，语法体系的复杂程度应该与该语言的发达程度有关，鉴于定量分析的困难，我们还需寻找更明确的标准。

3）词汇。随着语言的发展，词汇量的增加表现出稳定性和持续性的特点，比较客观地反映了一种语言的发达程度。一个词可以表达几个不同的概念（有多个义项），一个概念（义项）可以由多个不同的词来表达。语言产生之初，一个词对应一个概念；随着需要表达的概念的增多，语言发展的经济原则（简化规律）起作用，人们开始使用一个词表达多个概念；当对概念的认识进一步深化，语言的丰富化规律起作用，人们开始使用不同的词语表达细分的概念或新概念。概念与词之间的数量比例并不严格，只要保证词汇对人脑中概念体系的全覆盖，并可以充分区别不同概念间的微小差异，那么词汇量多一些少一些都是无关紧要的。同样的结论也适用于词素。

我们归纳出如下结论：音素的熵主要受生理因素和物理因素的影响，语言间和古今间的差异都不大；音节的熵主要受该语言音义结合特点的影响，具有民族性，语言间和古今间可以形成较大差异；音素熵和音节熵都与语言发达程度无关。语法系统表达词汇关系意义，它的复杂程度取决于词汇系统的复杂程度。词汇系统主要表达概念和概念间的关系，它的复杂程度取决于人脑中概念体系的复杂度。

图表 5　相同的概念体系 • 不同的语言表征

这样，我们就找到了衡量语言发达程度的标准——语言的词汇所表达的概念体系的复杂程度，或许可以称之为概念熵。这个结论是不难理解的：人类借助语言进行思维，思维的结果产生新思想新发现，借助语音系统把新思想新发现概念化成新词，充实进语言体系，词汇量的增加促进了词类之间关系的复杂化，于是调整或增加语法规则来表达复杂化了的词类关系，得到充实的语言体系推动思维活动进入更高的水平，于是完成了一个语言和思维相互促进的循环。在这个过程中，语言背后的概念体系是反映语言发达程度最客观的标准，而词汇量则是这个标准的外在表现形式。

在此基础上，我们还得到一个关于语言共性与个性的推论：在理想状态下，处于相同发达程度的语言，词汇所反映的概念系统的复杂程度相当；但是，这些概念通过不同语言的语音、词汇、语法手段表达出来，却表现出各不相同的外部形态和信息熵。这里要强调“理想状态”，因为现实中不存在发达程度恰好相同的语言；即使相同，经济社会等各领域的发展也会存在不平衡。该推论驳斥了依据语言的某些外部特点推断语言“先进性”的做法。图表 5 说明了语言间这种共性与个性的关系。（完）

参考文献

1）Shannon, C. E.「*Prediction and Entropy of Printed English*」『The BELL System Technical Journal』1951, Jan.
2）程工「语言共性的心理学和生理学证据」『解放军外国语学院学报』1999, (5).
3）冯志伟「汉字的熵」『语文建设』1984, (4).
4）冯志伟「汉字的信息量大不利于中文信息处理——再谈汉字的熵」『语文建设』1994, (3).
5）冯志伟「汉字和汉语的计算机处理」『当代语言学』2001, (1).
6）吕公礼「语用形式化与话语信息量研究」『外国语』2000, (6).
7）史磊，吕强「TrueType 字形描述技术和 TTF 文件」『中文信息』1995, (5).
8）王德春「论语义与认知」『外语电化教学』2009, (5).
9）王德春『语言学概论』上海：上海外语教育出版社　1997.
10）王晓龙，关毅『计算机自然语言处理』北京：清华大学出版社　2005.
11）徐通锵「汉语的特点和语言共性的研究」『语文研究』1999, (4).
12）徐扬「基于最大熵模型的汉语隐喻现象识别」『计算机工程与科学』2007, (4).

*　*　*　*　*　*

*　本文得到“上海外国语大学青年教师科研创新团队”项目资助。
*　上海外国语大学，国际文化交流学院，博士，讲师。